Significance Testing and Group Variable Selection

Adriano Zambom and Michael Akritas

The Pennsylvania State University

April 27, 2012

Abstract

Let $\mathbf{X}$, $\mathbf{Z}$ be r and s-dimensional covariates, respectively, used to model the response variable Y as $Y = m(\mathbf{X}, \mathbf{Z}) + \sigma(\mathbf{X}, \mathbf{Z})\epsilon$. We develop an ANOVA-type test for the null hypothesis that $\mathbf{Z}$ has no influence on the regression function, based on residuals obtained from local polynomial fitting of the null model. Using p-values from this test, a group variable selection method based on multiple testing ideas is proposed. Simulations studies suggest that the proposed test procedure outperforms the generalized likelihood ratio test when the alternative is non-additive or there is heteroscedasticity. Additional simulation studies, with data generated from linear, non-linear and logistic regression, reveal that the proposed group variable selection procedure performs competitively against Group Lasso, and outperforms it in selecting groups having nonlinear effects. The proposed group variable selection procedure is illustrated on a real data set.

Keywords: Nonparametric regression; local polynomial regression; Lack-of-fit tests; Dimension reduction; Backward elimination.

Acknowledgments: This research was partially supported by CAPES/Fulbright grant

[1]Adriano Zanin Zambom: adriano.zambom@gmail.com, Michael Akritas: mga@stat.psu.edu

15087657 and NSF grant DMS-0805598.

1 Introduction

Advances in data collection technologies and data storage devices have enabled the collection of data sets involving a large number of observations on many variables in several disciplines. When the objective of data collection is that of building a predictive model for a response variable, the challenges presented by massive data sets have opened new frontiers for statistical research. While the inclusion of a large number of predictors reduces modeling bias, the practice of including insignificant variables is likely to result in complicated models with less predictive power and reduced ability to discern and interpret the influence of the predictors. The underlying principles of modern model building are parsimony and sparseness. Parsimony requires simple models based on few predictors. Sparseness is a relatively new concept which evolved from the realization that in most scientific contexts prediction can be based on only a few variables. Variable selection uses the assumption of sparseness, enabling parsimonious model building. Thus, variable (also called feature) selection plays a central role in current scientific research as a fundamental component of model building.

Due to readily available software, variable selection is often performed by modeling the expected response at covariate value $\mathbf{x}$ as $m(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$. Classical approaches to variable selection, such as stepwise selection or elimination procedures, and best subset variable selection, can be computationally intensive or ignore stochastic errors. A new class of methodologies addresses variable selection through minimization of a constrained or penalized objective function, such as Tibshirani's (1996) LASSO, Fan and Li's (2001) SCAD, Efron, Hastie, Johnstone and Tibshirani's (2004) least angle regression, Zou's (2006) adaptive LASSO, and Candes and Tao's (2007) Dantzig selector. A different approach exploits the conceptual connection between model testing and variable selection: dropping variable j from the model is equivalent to not rejecting the null hypothesis $H_0^j : \beta_j = 0$. Abramovich, Benjamini, Donoho and Johnstone (2006) bridged the methodological divide by showing that the application

of the false discovery rate (FDR) controlling procedure of Benjamini and Hochberg (1995) on p-values resulting from testing each H_0^j can be translated into minimizing a model selection criterion similar to that used in Tibshirani and Knight (1999), Birge and Massart (2001) and Foster and Stine (2004). These criteria are more flexible than that of Donoho and Johnstone (1994), which uses a penalty parameter depending only on the dimensionality of the covariate, as well as AIC and Mallow's C_p, which use a constant penalty parameter. Working with orthogonal designs, Abramovich et al. (2006) showed that their method is asymptotically minimax for ℓ^r loss, $0 < r \leq 2$, simultaneously throughout a range of sparsity classes, provided the level q for the FDR is set to $q < 0.5$. Generalizations of this methodology to non-orthogonal designs differ mainly in the generation of the p-values for testing $H_0^j : \beta_j = 0$, and the FDR method employed. Bunea, Wegkamp and Auguste (2006) use p-values generated from the standardized regression coefficients resulting from fitting the full model and employ Benjamini and Yekuteli's (2001) method for controlling FDR under dependency, while Benjamini and Gavrilov (2009) use p-values from a forward selection procedure where the ith stage p-to-enter is the ith stage constant in the multiple-stage FDR procedure in Benjamini, Krieger and Yekutieli (2006).

Model checking and variable selection procedures based on the assumption of a linear model may fail to discern the relevance of covariates whose effect on $m(\mathbf{x})$ is nonlinear. Because of this, procedures for both model checking and variable selection have been developed under more general/flexible models. See, for example Li and Liang (2008), Wang and Xia (2008), Huang, Horowitz and Wei (2010), Storlie, Bondell, Reich and Zhang (2011), and references therein. However, the methodological approaches in this literature have been distinct from those of model checking. Working under a fully nonparametric regression model, Zambom and Akritas (2012) developed a competitive variable selection procedure by exploiting the aforementioned conceptual connection between model checking and variable selection. Their approach consists of backward elimination using the Benjamini and Yekuteli (2001)

method applied on the p-values resulting from testing the significance of each covariate. The test procedure they developed is based on the residuals obtained by fitting all covariates except the one whose significance is being tested. These residuals serve as the response variable in a one-way high-dimensional ANOVA design whose factor levels are the values of the covariate being tested. By augmenting these factor levels, and using smoothness assumptions, they developed an asymptotic theory for an ANOVA-type test statistic.

In many applications, covariates come in groups. For example, microarray experiments generate very large datasets with expression levels for thousands of genes but, typically, small sample size. Studies show that genes can act together as groups, and the scientific task is that of selecting the groups that are strongly associated with an outcome variable of interest. This type of problem can be addressed by first forming groups of genes through a clustering method and then selecting the important groups through a group selection procedure. One of the most common group selection procedures is the Group Lasso (Yuan and Lin, 2006), and the Adaptive Group Lasso (Wang and Xia, 2008). See also Park, Hastie and Tibshirani (2007) who, using averages of the genes within each group, perform a selection based on a procedure combining hierarchical clustering and Lasso.

The first part of is paper develops an extension of the ANOVA test procedure of Zambom and Akritas (2012) to testing the significance of a group of variables under a fully nonparametric model which also allows for heteroscedasticity. The second part of the paper introduces a backward elimination procedure for group variable selection using the Benjamini and Yekuteli (2001) method applied on the p-values resulting from testing the significance of each group.

This paper is organized as follows. Section 2 describes the proposed methodology for testing the significance of a group of variables, derives the asymptotic null distribution of the test statistic, and presents results of simulation studies comparing its performance to that of the generalized likelihood ratio test of Fan and Jiang (2005). Section 3 describes

the test-based group variable selection procedure, and presents results of simulation studies comparing its performance to that of Group Lasso. The analysis of a real data set involving gene expression levels of healthy and cancerous colon tissues is presented in Section 4.

2 Nonparametric Model Checking

2.1 The Hypothesis and the Test Statistic

Assume we have n observations, $(Y_i, \mathbf{U}_i)$, $i = 1, \ldots, n$, of the response variable Y and covariates $\mathbf{U} = (\mathbf{X}, \mathbf{Z})$, where $\mathbf{X}$ and $\mathbf{Z}$ have dimensions r and s respectively ($r + s = d$). Let $m(\mathbf{x}, \mathbf{z}) = E(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z})$ denote the regression function. The heterocedastic nonparametric regression model is

$$Y = m(\mathbf{X}, \mathbf{Z}) + \sigma(\mathbf{X}, \mathbf{Z})\epsilon, \tag{1}$$

where ϵ has zero mean and constant variance and is independent from $\mathbf{X}$ and $\mathbf{Z}$. The goal is to test the null hypothesis that $\mathbf{Z}$ does not contribute to the regression function, i.e.

$$H_0 : m(\mathbf{x}, \mathbf{z}) = m_1(\mathbf{x}). \tag{2}$$

The idea for testing this hypothesis is to treat the covariate values $\mathbf{Z}_i$, $i = 1, \ldots, n$, as the levels of high-dimensional one-way ANOVA design, with the null hypothesis residual $\hat{\xi}_i = Y_i - \hat{m}_1(\mathbf{x}_i)$ being the observation from factor level $\mathbf{Z}_i$, and construct an ANOVA based test statistic. Because the asymptotic theory for high-dimensional ANOVA requires more than one observation per factor level (Akritas and Papadatos, 2004), we will employ smoothness conditions, which will be stated below, and augment each factor level by including residuals from nearby covariate values. With a univariate covariate, such factor level augmentation was carried out in Wang, Akritas and Van Keilegom (2008) and Zambom and Akritas (2012) by ordering the covariate values and including in each factor level the residuals corresponding

to neighboring covariate values. With a multivariate covariate, the challenge is to order the factor levels, and hence the residuals, in a meaningful way resulting in a test statistic with good power properties. To do so, we propose to replace each $\mathbf{Z}_i$ by a nonlinear version of Bair, Hastie, Paul and Tibshirani's (2004) first supervised principal component (PC), $P_{\theta,i} = \mathbf{Z}_i^T \mathbf{C}_\theta$. The subscript θ will be explained below when the supervised PC is introduced.

Having a univariate surrogate of $\mathbf{Z}$, we augment each cell $P_{\theta,i} = \mathbf{Z}_i^T \mathbf{C}_\theta$ by including additional $p-1$, for p odd, residuals $\hat{\xi}_\ell$ which correspond to the $p-1$ nearest neighbors $P_{\theta,\ell}$ of $P_{\theta,i}$. To be specific, we consider the $(\hat{\xi}_i, P_{\theta,i})$, $i = 1, \ldots, n$, arranged so that $P_{\theta,i_1} < P_{\theta,i_2}$ whenever $i_1 < i_2$, and for each $P_{\theta,i}$, $(p-1)/2 < i \leq n-(p-1)/2$, define the nearest neighbor window W_i as

$$W_i(\mathbf{C}_\theta) = \left\{ j : |\hat{F}_P(P_{\theta,j}) - \hat{F}_P(P_{\theta,i})| \leq \frac{p-1}{2n} \right\}, \tag{3}$$

where $\hat{F}_P$ is the empirical distribution function of $P_{\theta,1}, \ldots, P_{\theta,n}$. $W_i(\mathbf{C}_\theta)$ defines the augmented cell corresponding to $P_{\theta,i}$. Note that the augmented cells are defined as sets of indices rather than as sets of $\hat{\xi}_i$ values. The vector of $(n-p+1)p$ constructed "observations" in the augmented one-way ANOVA design is

$$\hat{\boldsymbol{\xi}}_{\mathbf{C}_\theta} = (\hat{\xi}_j, j \in W_{(p-1)/2+1}(\mathbf{C}_\theta), \ldots, \hat{\xi}_j, j \in W_{n-(p-1)/2}(\mathbf{C}_\theta))^T. \tag{4}$$

Let MST and MSE denote the balanced one-way ANOVA mean squares due to treatment and error, respectively, computed on the data $\hat{\boldsymbol{\xi}}_{\mathbf{C}_\theta}$. The proposed test statistic is based on

$$MST - MSE. \tag{5}$$

In this paper the residuals $\hat{\xi}_i = Y_i - \hat{m}_1(\mathbf{x}_i)$, $i = 1, \ldots, n$, will be formed using the local polynomial of order q regression estimator defined by

$$\hat{m}_1(\mathbf{X}_i) = \mathbf{e}_1^T \left(\mathbb{X}_{\mathbf{X}_i}^T \mathbb{W}_{\mathbf{X}_i} \mathbb{X}_{\mathbf{X}_i} \right)^{-1} \mathbb{X}_{\mathbf{X}_i}^T \mathbb{W}_{\mathbf{X}_i} \mathbf{Y} = \sum_{j=1}^{n} \tilde{w}(\mathbf{X}_i, \mathbf{X}_j) Y_j, \ i = 1, \ldots, n, \tag{6}$$

where $\mathbb{W}_{\mathbf{x}} = diag\{K_{H_n}(\mathbf{X}_1 - \mathbf{x}), \ldots, K_{H_n}(\mathbf{X}_n - \mathbf{x})\}$, with $K_{H_n}(\mathbf{x}) = |H_n|^{-1/2} K(H_n^{-1/2}\mathbf{x})$ for $K(\cdot)$ a bounded, non-negative r-variate kernel function of bounded variation and with bounded support and $H_n^{1/2}$ is a symmetric positive definite $r \times r$ bandwidth matrix, and

$$\mathbb{X}_{\mathbf{x}} = \begin{pmatrix} 1 & (\mathbf{X}_1 - \mathbf{x})^T & \text{vech}^T\left\{(\mathbf{X}_1 - \mathbf{x})(\mathbf{X}_1 - \mathbf{x})^T\right\} & \ldots \\ \vdots & \vdots & \vdots & \ldots \\ 1 & (\mathbf{X}_n - \mathbf{x})^T & \text{vech}^T\left\{(\mathbf{X}_n - \mathbf{x})(\mathbf{X}_n - \mathbf{x})^T\right\} & \ldots \end{pmatrix},$$

with vech denoting the half-vectorization operator, is the $n \times \gamma_{r,q}$ design matrix, where

$$\gamma_{r,q} = \sum_{j=0}^{q} \sum_{\substack{k_1=0 \\ k_1+\ldots+k_r=j}}^{j} \ldots \sum_{k_r=0}^{j} 1.$$

We finish this section with a description of the construction of the first non-linearly supervised principal component $P_\theta = \mathbf{Z}^T \mathbf{C}_\theta$. Let $p_j, j = 1, \ldots, s$, denote the p-values obtained by applying the test of Zambom and Akritas (2012) for testing the hypothesis H_0^j which specifies that Z_j, the jth coordinate of $\mathbf{Z}$, has no effect on the regression function of the model with response variable Y and covariate vector $(\mathbf{X}, Z_j)$. For a threshold parameter θ, define the index set $\mathcal{J} = \{j : p_j < \theta\}$ and let $\mathbf{Z}_{\mathcal{J}}$ be the vector formed from the $\mathcal{J}$ coordinates of $\mathbf{Z}$. Then, $P_\theta = \mathbf{Z}^T \mathbf{C}_\theta$ is the first principal component of $\mathbf{Z}_{\mathcal{J}}$. Note that some entries of $\mathbf{C}_\theta$ are equal to 0, corresponding to the coordinates of $\mathbf{Z}$ with p_j greater or equal to θ. It is important to keep in mind that the observable vector of first nonlinear principal components, $\mathbf{P}_\theta = (P_{\theta,1}, \ldots, P_{\theta,n})$, depends on the estimated residuals $\hat{\xi}_i, i = 1, \ldots, n$, to the extend that $\mathcal{J}$, and hence $\mathbf{C}_\theta$, depend on them.

2.2 Asymptotic null distribution

Theorem 2.1. *Assume that the marginal densities $f_{\mathbf{X}}$, $f_{\mathbf{Z}}$ of $\mathbf{X}$, $\mathbf{Z}$, respectively, are bounded away from zero, the $q+1$ derivatives of $m_1(\mathbf{x})$ are uniformly continuous and bounded, that $\sigma^2(., \mathbf{z}) := E(\xi^2 | \mathbf{Z}^T \mathbf{C})$ is Lipschitz continuous, $\sup_{\mathbf{x},\mathbf{z}} \sigma^2(\mathbf{x}, \mathbf{z}) < \infty$, and $E(\epsilon_i^4) < \infty$. Assume*

that the eigenvalues, $\lambda_i,\ i = 1, \ldots, r$*, of the bandwidth matrix* $H_n^{1/2}$ *converge to zero at the same rate and satisfy*

$$n\lambda_i^{4(q+1)} \to 0 \quad \textit{and} \quad \frac{n\lambda_i^{2r}}{(\log n)^2} \to \infty,\ i = 1, \ldots, r. \tag{7}$$

Then, under H_0 *in (2), the asymptotic distribution of the test statistic in (5) is given by*

$$n^{1/2}(MST - MSE) \xrightarrow{d} N(0, \frac{2p(2p-1)}{3(p-1)}\tau^2),$$

where $\tau = \int \left[\int \sigma^2(\mathbf{x}, \mathbf{z}) f_{\mathbf{X}|\mathbf{Z}^T\mathbf{C}=\mathbf{z}^T\mathbf{C}}(\mathbf{x}) d\mathbf{x}\right]^2 f_{\mathbf{Z}^T\mathbf{C}}(\mathbf{z}^T\mathbf{C}) d(\mathbf{z}^T\mathbf{C})$.

An estimate of τ^2 can be obtained by modifying Rice's (1984) estimator as follows

$$\hat{\tau}^2 = \frac{1}{4(n-3)} \sum_{j=2}^{n-2} (\hat{\xi}_j - \hat{\xi}_{j-1})^2 (\hat{\xi}_{j+2} - \hat{\xi}_{j+1})^2. \tag{8}$$

Asymptotic theory under local additives and under general local alternatives is derived in Zambom (2012). As these limiting results show, the asymptotic mean of the test statistic $MST - MSE$ is positive under alternatives. Thus, the test procedure rejects the null hypothesis for "large" values of the test statistic.

2.3 Simulations: Model Checking Procedures

We compare the proposed ANOVA-type hypothesis test for groups with the generalized likelihood ratio test of Fan and Jiang (2005). The data is generated under three situations: a homoscedastic additive model, a homoscedastic non-additive model, and a heterocedastic non-additive model. All covariates, in all models, are independent standard normal. The homoscedastic additive model is

$$Y = X_1 + \theta(Z_1 + Z_2 + Z_3) + \epsilon, \quad \text{where} \quad \epsilon \sim N(0, 1), \tag{9}$$

the homoscedastic non-additive model is

$$Y = X_1^{X_2}(1 + \theta(Z_1 + Z_2)) + X_2^{\theta(Z_1+Z_2)} + \epsilon, \quad \text{where} \quad \epsilon \sim N(0, .1^2), \tag{10}$$

and the heterocedastic non-additive model is

$$Y = X_1 + \theta \sin(Z_1 Z_2) + Z_1 Z_2 \epsilon, \quad \text{where} \quad \epsilon \sim N(0, .5^2). \tag{11}$$

In each situation we simulate 2000 data sets ot size $n = 200$. All simulations were performed in R.

In order to evaluate the effect of the threshold parameter θ we applied our test procedure with $\theta = 0.05$ and $\theta = 0.2$. Moreover, in each case we considered two rules to form the set of covariates from which the first supervised principal component is obtained. Rule 1 consists of using only the covariates with p-value less than θ, and in Rule 2 we consider the set of covariates chosen from Rule 1 and add to the set the covariate with the smallest p-value among the remainder covariates. In each case, if the number of selected covariates is less than two the set is formed from the two with the smallest p-value. Thus the simulations consider four versions of our test statistic: a) Rule 1 with $\theta = 0.05$, b) Rule 1 with $\theta = 0.2$, c) Rule 2 with $\theta = 0.05$, d) Rule 2 with $\theta = 0.2$. All four versions of our test statistic use windows of $p = 11$.

Tables 1, 2, and 3, show the simulation results for models (9),(10), and (11), respectively. It is seen that the proposed test procedure is robust to the choice of the threshold parameter, and to the rules for selecting the set of covariates from which the first supervised principal component is obtained. The Generalized Likelihood Ratio test, which is designed for homoscedastic additive models, achieves better power under model (9), but is extremely liberal under heteroscedasticity and its power for the non-additive alternatives of model (11) is mainly less than its level; see Table 3. Table 2 suggests that the GRLT has low power against non-additive alternatives even in the homoscedastic case.

Table 1: Rejection rates for the homocedastic additive model

	θ				
Method	0	.2	.4	.6	.8
ANOVA-type-a	.066	.404	.691	.706	.751
ANOVA-type-b	.060	.378	.613	.689	.733
ANOVA-type-c	.066	.396	.600	.692	.749
ANOVA-type-d	.057	.375	.618	.685	.718
GRLT	.048	.883	1	1	1

Table 2: Rejection rates for the homocedastic non-additive model

	θ				
Method	0	.02	.04	.06	.08
ANOVA-type-a	.051	.202	.522	.693	.724
ANOVA-type-b	.047	.192	.560	.710	.739
ANOVA-type-c	.050	.193	.520	.679	.711
ANOVA-type-d	.047	.161	.510	.676	.733
GRLT	.052	.059	.117	.235	.379

3 Nonparametric Group Variable Selection

In this section we will present a test-based group variable selection. For this purpose we will make a slight change in notation by letting $\mathbf{X}$ denote the entire vector of covariates. Thus, we consider the nonparametric regression model

$$Y_i = m(\mathbf{X}_i) + \sigma(\mathbf{X}_i)\varepsilon_i,\ i = 1, \ldots, n, \tag{12}$$

where ε_i is the independent error with zero mean and constant variance. Suppose that the covariates are classified in d groups identified by the indices $J_\ell = \{j : X_j \text{ belongs to group } \ell\}$, $\ell = 1, \ldots, d$, and let s_ℓ denote the size of group J_ℓ. Moreover, we assume sparseness in the sense that only the variables in a subset $I_0 = \{J_1, \ldots, J_{d_0}\} \subset \{J_1, \ldots, J_d\}$ of the groups influence the regression function. Finally, we will assume the dimension reduction model of Li (1991), i.e.

$$m(\mathbf{x}) = g(\mathbf{B}\mathbf{x}),\ \text{where } \mathbf{B} \text{ is a } K \times (\textstyle\sum_i s_i) \text{ matrix.} \tag{13}$$

Table 3: Rejection rates for the heterocedastic non-additive model

	θ				
Method	0	.3	.6	1	2
ANOVA-type-a	.035	.168	.503	.654	.789
ANOVA-type-b	.040	.172	.529	.663	.767
ANOVA-type-c	.037	.161	.501	.651	.757
ANOVA-type-d	.036	.190	.520	.657	.742
GRLT	.584	.585	.535	.439	.297

Define the hypothesis

$$H_0^\ell : m(\mathbf{x}) = m_1(\mathbf{x}_{(-J_\ell)}),\ \ell = 1, \ldots, d,$$

where $\mathbf{x}_{(-J_\ell)}$ is the set of all covariates except those whose index are in J_ℓ. Under the dimension reduction model (13), this hypothesis can be written equivalently as

$$H_0^\ell : g(\mathbf{Bx}) = g(\mathbf{B}_{(-J_\ell)}\mathbf{x}_{(-J_\ell)}),\ \ell = 1, \ldots, d, \tag{14}$$

where $\mathbf{B}_{(-J_\ell)}$ is the $K \times (d - s_\ell)$ matrix obtained by omitting the columns of $\mathbf{B}$ with indices in J_ℓ. Let $\widehat{\mathbf{B}}$ denote the Sliced Inverse Regression (SIR) estimator of $\mathbf{B}$, and $\widehat{\mathbf{B}}_{(-J_\ell)}$ be the corresponding submatrix. With this notation, let

$$z_\ell = \sqrt{n}(MST_\ell - MSE_\ell)/\sqrt{\frac{2p(2p-1)}{3(p-1)}\hat{\tau}_\ell^4}$$

be the test statistic for testing the hypothesis (14) with $\widehat{\mathbf{B}}_{(-J_\ell)}\mathbf{X}_{(-J_\ell)}$ playing the role of $\mathbf{X}$ in Theorem 2.1, and $\widehat{\mathbf{B}}_{(J_\ell)}\mathbf{X}_{(J_\ell)}$ playing the role of $\mathbf{Z}$, where $\mathbf{X}_{(J_\ell)}$ is the set of all covariates whose index are in J_ℓ and $\widehat{\mathbf{B}}_{(J_\ell)}$ is the corresponding submatrix of $\widehat{\mathbf{B}}$.

In this context we will describe the following group variable selection procedure using backward elimination based on the Benjamini and Yekuteli (2001) method for controlling the false discovery rate (FDR):

1. Compute the p-value for H_0^ℓ as $\pi_\ell = 1 - \Phi(z_\ell)$, $\ell = 1, \ldots, d$.

2. Compute

$$k = \max \left\{ i : \pi_{(\ell)} \leq \frac{\ell}{d} \frac{\alpha}{\sum_{j=1}^{d} j^{-1}} \right\} \tag{15}$$

for a choice of level α, where $\pi_{(1)}, \ldots, \pi_{(d)}$ are the ordered p-values. If $k = d$ stop and retain all groups. If $k < d$

(a) update $\mathbf{x}$ by eliminating the covariates of the group corresponding to $\pi_{(d)}$,

(b) update d to $d - 1$,

(c) update $\widehat{\mathbf{B}}$ by eliminating the columns corresponding to the deleted variables,

(d) update the test statistic z_ℓ, $\ell = 1, \ldots, d$.

3. Repeat steps 1 and 2, with the updated z_ℓ, $\ell = 1, \ldots, d$.

Remark Another approach for constructing a group variable selection procedure is to use a single application of the Benjamini and Yekuteli (2001) method for controlling the false discovery rate (FDR). This is similar to one of the two procedures proposed in Bunea et al. (2006). However, this did not perform well in simulations and is not recommended. A backward elimination approach was used in Li, Cook and Nachtsheim (2005), but without incorporating multiple testing ideas.

3.1 Simulations: Variable selection procedure

In this section we compare the variable selection based on the ANOVA-type test to the Group Lasso proposed by Yuan and Lin (2006). We study the behavior of the selection for two different scenarios, one with a continuous response and another with a binary response.

For the continuous response scenario the data is generated according to the models

$$\begin{aligned}
\text{Model 1}: Y &= X_3^3 + X_3^2 + X_3 + (1/3)X_6^3 - X_6^2 + (2/3)X_6 + \epsilon \\
\text{Model 2}: Y &= sin(X_3^3 + X_3^2 + X_3) + (1/3)X_6^3 - X_6^2 + (2/3)X_6 + \epsilon \\
\text{Model 3}: Y &= 10sin(X_3^3 + X_3^2 + X_3) + 5sin((1/3)X_6^3 - X_6^2 + (2/3)X_6) + \epsilon
\end{aligned}$$

where $X_i = (Z_i + W)/\sqrt{2}$, $Z_i, i = 1, \ldots, 16$ and W iid $N(0,1)$, and $\epsilon \sim N(0, 2^2)$. Thus, for Models 1, 2 and 3 there are 16 groups of three covariates each, represented by the polynomial terms. The only groups that are significant are groups 3 and 6. We run 1000 simulations of data sets of size $n = 100$. Table 4 shows the mean number of correct and incorrect groups selected by the ANOVA-type variable selection and Group Lasso using the C_p criterion. It is seen that Group Lasso tends to select more groups that are not significant to the regression, while both methods perform competitively in selecting the significant groups.

Table 4: Results for the ANOVA-type and Group Lasso

Model	Method	Corr.Selected	Incorr.Selected
Model 1	ANOVA-type	1.80	.55
	Group LASSO	2	4.7
Model 2	ANOVA-type	1.15	.81
	Group LASSO	1.59	4.21
Model 3	ANOVA-type	1.84	0.64
	Group LASSO	1.80	6.75

For the second scenario, we consider the following three logistic regression models.

$$\text{Models 1 and 2}: \quad p_j(\mathbf{X}) = \frac{1}{1 + exp(-\boldsymbol{\beta}_j^T(1, \mathbf{X})^T)}, \ j = 1, 2,$$

where $\mathbf{X} = (X_1, \ldots, X_{15})$ are iid $U(0,1)$, grouped sequentially in 5 groups of 3 covariates each, and

$$\begin{aligned}
\boldsymbol{\beta}_1 &= (1, -2.2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 0, 0, 0)^T \\
\boldsymbol{\beta}_2 &= (1, -2.2, 3, 0, 0, 0, 0, 0, 0, 0, 0, 1, 3, 0, 0, 0, 0)^T.
\end{aligned}$$

$$\text{Model 3}: \quad p_3(\mathbf{X}) = \frac{1}{1 + exp\left(-18\sin(\pi X_2) - 18\sin(\pi X_8)\right)},$$

Table 5: Results for logistic regression

Model	Method	Corr.Selected	Incorr.Selected
Model 1	ANOVA-type	.340	.261
	Group LASSO	.197	.032
Model 2	ANOVA-type	.287	.312
	Group LASSO	.100	.021
Model 3	ANOVA-type	1.223	0.080
	Group LASSO	0.040	0.039

where $\mathbf{X} = (X_1, \ldots, X_{12})$, with $X_1, \ldots, X_{11}$ iid $U(0, 3)$ and $X_{12} \sim N(-3, 1)$ independent of the others, are grouped sequentially in 4 groups of 3 covariates each.

The results in Table 5 are based on 1000 simulation runs using $n = 100$ for Models 1 and 2, and $n = 200$ for Model 3. It is seen that for Models 1 and 2 the number of correctly selected covariates by either procedure is low. This is probably due to the smaller sample size and the larger number of covariates. For Model 3, the Group Lasso fails to select covariates, while the ANOVA-type procedure seems to perform very well. In summary, the simulation results suggest that the ANOVA-type variable selection procedure outperforms the Group Lasso when the logistic regression model involves a non-linear function of the covariates, and has competitive performance in the other cases.

4 Real Data Example

The proposed procedure will be illustrated with an analysis of the colon cancer dataset of Alon et al. (1999). The dataset was obtained from the Affymetrix technology and shows expression levels of 40 tumor and 22 normal colon tissues of 6,500 human genes. A selection of 2,000 genes with highest minimal intensity across the samples has been made by Alon et al. (1999) and is publicly available at http://microarray.princeton.edu/oncology. Different clustering methods have been applied to this data set in several previous studies including Dettling and Buhlman (2002, 2004), and Ma, Song and Huang (2007).

To illustrate the proposed ANOVA-type group variable selection procedure we first apply a clustering method to form the groups. We chose the supervised clustering procedure *Wilma* proposed by Dettling and Buhlman (2002) which is available in the package supclust in R. *Wilma* requires as input the number of clusters to be formed, and we specified 60, 55, 50 and 45 clusters. The next step of the proposed procedure requires dimension reduction through SIR. However, because the number of genes (2,000) is much larger than the sample size (62) it is not possible to use SIR straightforward. Therefore to estimate $\mathbf{B}$, we ran SIR on the set of predictors composed of the first supervised principal component of each cluster.

We also ran the Group Lasso procedure for binary responses using the package grplasso in R on the same clusters/groups returned by Wilma. However, a corresponding modification is needed for the calculation of the degrees of freedom needed for the application of the C_p criterion; see Yuan and Lin (2006). This calculation requires the estimator $\hat{\beta}$ from fitting all individual covariates. Since the number of covariates is much larger than the sample size, we obtain an approximation to the required estimator by first obtaining the estimator $\hat{\beta}_P$ from fitting the first PC from each cluster. Since the PCs are linear combinations of the covariates in each cluster, having a coefficient for a cluster's PC translates into coefficients for the covariates in that cluster.

Table 6: Results for Colon data set

No. Initial Clusters	Procedure	Clusters Selected
60	ANOVA-type	1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 15, 16, 18, 23, 24, 25, 28, 31, 33, 34, 39, 42, 46
	Group Lasso	1, 2, 3, 8, 9, 16, 17, 37
55	ANOVA-type	1, 2, 3, 4, 5, 6, 8, 10, 11, 12, 15, 16, 17, 22, 23, 24, 25, 26, 33, 38, 42, 45
	Group Lasso	2, 3, 7, 9, 13, 18
50	ANOVA-type	1, 2, 3, 4, 5, 6, 8, 10, 11, 12, 14, 15, 16, 18, 21, 23, 24, 25, 26, 36, 40, 45, 47, 48, 49
	Group Lasso	2, 3, 7, 9, 27, 29
45	ANOVA-type	1, 2, 3, 4, 5, 6, 8, 10, 11, 12, 15, 16, 18, 21, 22, 23, 24, 25, 31, 34, 37, 40, 42, 44, 45
	Group Lasso	2, 3, 7, 9, 10, 16

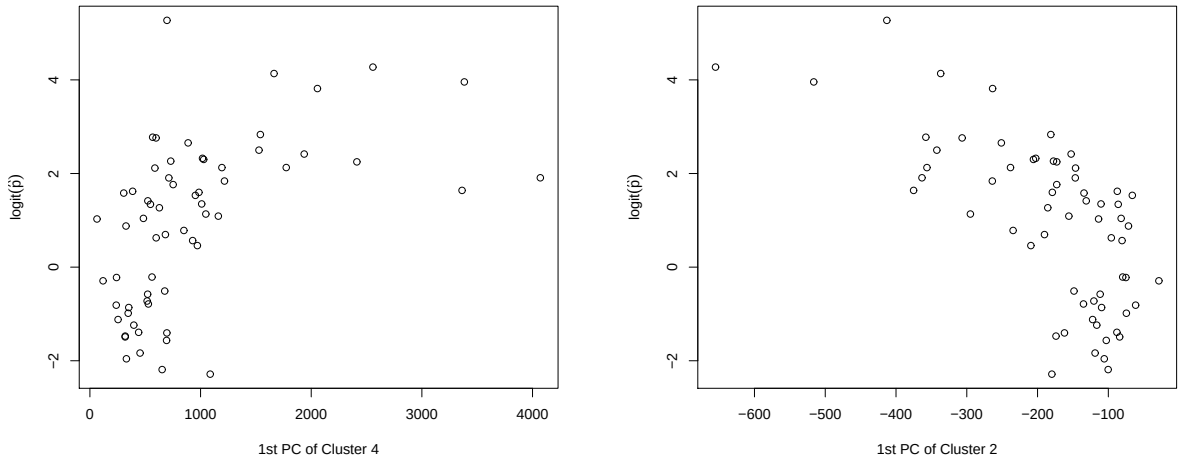

Figure 1: Plot of Group Lasso Logit Estimate Against the First PC of Clusters 4 and 2

Table 6 shows the groups selected by the proposed group variable selection and the group lasso procedure for the different specified number of clusters returned by Wilma.

We note that there is significant overlap in the genes included in the clusters selected by each method across the different numbers of total clusters specified. Thus, the different number of total clusters specified is not critical for selecting the important genes. The proposed method selects more clusters, which is contrary to the simulation results. This is probably due to the linear link function for the logit used in grplasso. For example, when the logit of the fitted probability of cancerous tissue is plotted against the first PC of cluster 4, which is selected by the proposed method but not by Group Lasso, it shows a nonlinear effect (left panel of Figure 1), whereas the non-linearity is much less pronounced when plotted against the first PC of cluster 2, which is selected by the proposed method and by Group Lasso (right panel of Figure 1).

A Appendix

Proof of Theorem 2.1. Under H_0 in (2) we write

$$\begin{aligned}\hat{\xi}_i &= Y_i - \hat{m}_1(\mathbf{X}_i) + m_1(\mathbf{X}_i) - m_1(\mathbf{X}_i) = \xi_i - (\hat{m}_1(\mathbf{X}_i) - m_1(\mathbf{X}_i)) \\ &= \xi_i - \Delta_{m_1}(\mathbf{X}_i),\end{aligned}$$

where $\Delta_{m_1}(\mathbf{X}_i)$ is defined implicitly in the above relation. Thus, $\hat{\boldsymbol{\xi}}_{\mathbf{C}_\theta}$ of relation (4) is decomposed as $\hat{\boldsymbol{\xi}}_{\mathbf{C}_\theta} = \boldsymbol{\xi}_{\mathbf{C}_\theta} - \boldsymbol{\Delta}_{m_1\mathbf{C}_\theta}$, where $\boldsymbol{\xi}_{\mathbf{C}_\theta}$ and $\boldsymbol{\Delta}_{m_1\mathbf{C}_\theta}$ are defined as in (4) but using ξ_i and $\Delta_{m_1}(\mathbf{X}_i)$, respectively, instead of $\hat{\xi}_i$. Note that MST-MSE given in (5) can be written as a quadratic form $\hat{\boldsymbol{\xi}}_{\mathbf{C}_\theta}^T A \hat{\boldsymbol{\xi}}_{\mathbf{C}_\theta}$ (see Wang, Akritas and Van Keilegom, 2008), where

$$A = \frac{np-1}{n(n-1)p(p-1)} \oplus_{i=1}^{n} \mathbf{J}_p - \frac{1}{n(n-1)p}\mathbf{J}_{np} - \frac{1}{n(p-1)}\mathbf{I}_{np}, \tag{16}$$

$\mathbf{I}_d$ is a identity matrix of dimension d, $\mathbf{J}_d$ is a dxd matrix of 1's and $\oplus$ is the Kronecker sum or direct sum. Thus, we can write $\sqrt{n}$(MST - MSE) as

$$\sqrt{n}\hat{\boldsymbol{\xi}}_{\mathbf{C}_\theta}^T A \hat{\boldsymbol{\xi}}_{\mathbf{C}_\theta} = \sqrt{n}\boldsymbol{\xi}_{\mathbf{C}_\theta}^T A \boldsymbol{\xi}_{\mathbf{C}_\theta} - \sqrt{n}2\boldsymbol{\xi}_{\mathbf{C}_\theta}^T A \boldsymbol{\Delta}_{m_1\mathbf{C}_\theta} + \sqrt{n}\boldsymbol{\Delta}_{m_1\mathbf{C}_\theta}^T A \boldsymbol{\Delta}_{m_1\mathbf{C}_\theta}. \tag{17}$$

That $\sqrt{n}2\boldsymbol{\xi}_{\mathbf{C}_\theta}^T A \boldsymbol{\Delta}_{m_1\mathbf{C}_\theta}$ and $\sqrt{n}\boldsymbol{\Delta}_{m_1\mathbf{C}_\theta}^T A \boldsymbol{\Delta}_{m_1\mathbf{C}_\theta}$ converge in probability to 0 uniformly follows from arguments similar to those used in Zambom and Akritas (2012).

Using Corolary A.1, to show the asymptotic normality of $\sqrt{n}\boldsymbol{\xi}_{\mathbf{C}_\theta}^T A \boldsymbol{\xi}_{\mathbf{C}_\theta}$, it is enough to show that

$$\sup_{\mathbf{C}} \left| P\left(\frac{\sqrt{n}\boldsymbol{\xi}_{\mathbf{C}}^T A_d \boldsymbol{\xi}_{\mathbf{C}}}{\frac{2p(2p-1)}{3(p-1)}\tau^2} \leq t \right) - \Phi(t) \right| \to 0.$$

Let $b_n \sim n^{2/3}$ and $r_n \sim n/b_n \sim n^{1/3}$ and write

$$\begin{aligned}\sqrt{n}\boldsymbol{\xi}_{\mathbf{C}_\theta}^T A_d \boldsymbol{\xi}_{\mathbf{C}_\theta} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{p-1} \sum_{j_1 \neq j_2} \xi_{j_1}\xi_{j_2} I\left(j_1, j_2 \in W_i(\mathbf{C}_\theta)\right) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{r_n} U_i(\mathbf{C}_\theta) + \frac{1}{\sqrt{n}} \sum_{i=1}^{r_n} V_i(\mathbf{C}_\theta) \\ &= \frac{1}{\sqrt{n}} S_U(C_\theta) + \frac{1}{\sqrt{n}} S_V(C_\theta),\end{aligned} \tag{18}$$

where, with $\gamma_i(\mathbf{C}_\theta) = \frac{1}{p-1}\sum_{j_1\neq j_2}\xi_{j_1}\xi_{j_2}I\left(j_1,j_2\in W_i(\mathbf{C}_\theta)\right)$,

$$\begin{aligned} U_i(\mathbf{C}_\theta) &= \gamma_{(i-1)(b_n+p)+1}(\mathbf{C}_\theta)+\ldots+\gamma_{(i-1)(b_n+p)+b_n}(\mathbf{C}_\theta), \\ V_i(\mathbf{C}_\theta) &= \gamma_{(i-1)(b_n+p)+b_n+1}(\mathbf{C}_\theta)+\ldots+\gamma_{i(b_n+p)}(\mathbf{C}_\theta). \end{aligned}$$

Note that the $U_i(\mathbf{C}_\theta)$ are independent, and the $V_i(\mathbf{C}_\theta)$ are independent.

Now, letting $\mathrm{sd}=\sqrt{\frac{2p(2p-1)}{3(p-1)}\tau^2}$, we have

$$\begin{aligned} &\sup_{\mathbf{C}}\left|P\left(\frac{\sqrt{n}\boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}}{\mathrm{sd}}\le t\right)-\Phi(t)\right| \\ =\ &\sup_{\mathbf{C}}\left|P\left(\frac{S_U(\mathbf{C})+S_V(\mathbf{C})}{\sqrt{n}\mathrm{sd}}\le t\right)-\Phi(t)\right| \\ =\ &\sup_{\mathbf{C}}\left|P\left(\frac{S_U(\mathbf{C})}{\sqrt{n}\mathrm{sd}}\le t-\frac{S_V(\mathbf{C})}{\sqrt{n}\mathrm{sd}},\left|\frac{S_V(\mathbf{C})}{\sqrt{n}\mathrm{sd}}\right|\le\epsilon\right)\right. \\ &\qquad\left.+P\left(\frac{S_U(\mathbf{C})}{\sqrt{n}\mathrm{sd}}\le t-\frac{S_V(\mathbf{C})}{\sqrt{n}\mathrm{sd}},\left|\frac{S_V(\mathbf{C})}{\sqrt{n}\mathrm{sd}}\right|\ge\epsilon\right)-\Phi(t)\right| \\ \le\ &\sup_{\mathbf{C}}\left|P\left(\frac{S_U(\mathbf{C})}{\sqrt{n}\mathrm{sd}}\le t-\frac{S_V(\mathbf{C})}{\sqrt{n}\mathrm{sd}},\left|\frac{S_V(\mathbf{C})}{\sqrt{n}\mathrm{sd}}\right|\le\epsilon\right)-\Phi(t)\right| \\ &+\sup_{\mathbf{C}}P\left(\left|\frac{S_V(\mathbf{C})}{\sqrt{n}\mathrm{sd}}\right|\ge\epsilon\right) \end{aligned} \tag{19}$$

That the second in (19) term converges to zero follows from Lemma A.2. That the first term in (19) converges to zero follows from Lemma A.3, provided we show that

$$\mathrm{Var}\left(\frac{S_U(\mathbf{C})}{\sqrt{n}}\right)\to\mathrm{sd}^2,\quad\text{for any}\quad\mathbf{C}. \tag{20}$$

By (18), and because $\frac{S_V(\mathbf{C})}{\sqrt{n}\mathrm{sd}}\xrightarrow{p}0$, (20) follows from $\sup_{\mathbf{C}}\mathrm{Var}(\sqrt{n}\boldsymbol{\xi}_{vC}^T A_d\boldsymbol{\xi}_{vC})\to\mathrm{sd}^2$. By the definition of $\boldsymbol{\xi}_{\mathbf{C}_\theta}^T A_d\boldsymbol{\xi}_{\mathbf{C}_\theta}$, it is easy to see that $E\left(\boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}\right)=0$ for any $\mathbf{C}$. To find the variance of $\sqrt{n}\boldsymbol{\xi}_{vC}^T A_d\boldsymbol{\xi}_{vC}$ we first evaluate the conditional second moment $E[(\sqrt{n}\boldsymbol{\xi}_{vC}^T A_d\boldsymbol{\xi}_{vC})^2|\mathbf{Z}^T\mathbf{C}]$.

Recalling the notation $\sigma^2(.,\mathbf{z}_j^T\mathbf{C}) = E(\xi_j^2|\mathbf{Z}^T\mathbf{C} = \mathbf{z}_j^T\mathbf{C})$, we have

$$\begin{aligned}
&\sup_{\mathbf{C}} \frac{1}{n(p-1)^2} \sum_{i_1,i_2}^{n} \sum_{j_1\neq l_1}^{n} \sum_{j_2\neq l_2}^{n} E(\xi_{j_1}\xi_{l_1}\xi_{j_2}\xi_{l_2}|\mathbf{Z}^T\mathbf{C}) I(j_s \in W_{i_s}(\mathbf{C}), l_s \in W_{i_s}(\mathbf{C}), s=1,2) \\
&= \sup_{\mathbf{C}} \frac{2}{n(p-1)^2} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \sum_{j\neq l}^{n} \sigma^2(.,\mathbf{z}_j^T\mathbf{C})\sigma^2(.,\mathbf{z}_l^T\mathbf{C}) I(j,l \in W_{i_1}(\mathbf{C}) \cap W_{i_2}(\mathbf{C})) \\
&= \sup_{\mathbf{C}} \frac{2}{n(p-1)^2} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \sum_{j\neq l}^{n} \sigma^2(.,\mathbf{z}_j^T\mathbf{C}) \left(\sigma^2(.,\mathbf{z}_j^T\mathbf{C}) + O_p\left(\frac{p}{\sqrt{n}}\right)\right) I(j,l \in W_{i_1}(\mathbf{C}) \cap W_{i_2}(\mathbf{C})) \\
&= \sup_{\mathbf{C}} \frac{2}{n(p-1)^2} \sum_{j=1}^{n} \sigma^4(.,\mathbf{z}_j^T\mathbf{C}) \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \sum_{l\neq j}^{n} I(j,l \in W_{i_1}(\mathbf{C}) \cap W_{i_2}(\mathbf{C})) + O_p\left(\frac{p^2}{n^{1/2}}\right) \\
&= \sup_{\mathbf{C}} \frac{2}{n(p-1)^2} \sum_{j=1}^{n} \sigma^4(.,\mathbf{z}_j^T\mathbf{C}) 2(1+2^2+3^2+...+(p-1)^2) + O_p\left(\frac{p^2}{n^{1/2}}\right) \\
&= \sup_{\mathbf{C}} \frac{2}{n(p-1)^2} \frac{p(p-1)(2p-1)}{3} \sum_{j=1}^{n} \sigma^4(.,\mathbf{z}_j^T\mathbf{C}) + O_p\left(\frac{p^2}{n^{1/2}}\right),
\end{aligned}$$

where the third equality follows from Lemma A.5 using the assumption that $\sigma^2(.,\mathbf{z}_j^T\mathbf{C})$ is Lipschitz continuous and the second last inequality results from the fact that if $1 \leq |j_1 - j_2| = s \leq p-1$, then they are $(p-s)^2$ pairs of windows whose intersection includes j_1 and j_2. Taking limits as $n \to \infty$ it is seen that

$$\sup_{\mathbf{C}} E\left(n^{1/2}\boldsymbol{\xi}_{\mathbf{C}}^T A_d \boldsymbol{\xi}_{\mathbf{C}} \middle| \mathbf{Z}^T\mathbf{C}\right)^2 \xrightarrow{a.s.} \frac{2(2p-1)}{3(p-1)} E(\sigma^4(.,\mathbf{z}^T\mathbf{C})) = \frac{2(2p-1)}{3(p-1)}\tau^2. \tag{21}$$

From relation (21) it is easily seen that $\sup_{\mathbf{C}} E[(\sqrt{n}\boldsymbol{\xi}_{\mathbf{C}}^T A_d \boldsymbol{\xi}_{\mathbf{C}})^2|\mathbf{Z}^T\mathbf{C}]$ remains bounded, and thus $\sup_{\mathbf{C}} \mathrm{Var}(n^{1/2}\xi_{\mathbf{C}}^T A \xi_{\mathbf{C}})$ also converges to the same limit by the Dominated Convergence Theorem. □

Lemma A.1. *If the assumptions of Theorem 2.1 hold, then under H_0 and as $n \to \infty$,*

$$\sup_{\mathbf{C}} P\left(n^{1/2}\left|\boldsymbol{\xi}_{\mathbf{C}_\theta}^T A \boldsymbol{\xi}_{\mathbf{C}_\theta} - \boldsymbol{\xi}_{\mathbf{C}_\theta}^T A_d \boldsymbol{\xi}_{\mathbf{C}_\theta}\right| \geq \epsilon\right) \to 0, \tag{22}$$

where $A_d = diag\{B_1,...,B_n\}$, with $B_i = \frac{1}{n(p-1)}[\mathbf{J}_p - \mathbf{I}_p]$.

Proof. By Chebyshev Inequality, we have that

$$\sup_{\mathbf{C}} P\left(n^{1/2}\left|\boldsymbol{\xi}_{\mathbf{C}}^T A \boldsymbol{\xi}_{\mathbf{C}} - \boldsymbol{\xi}_{\mathbf{C}}^T A_d \boldsymbol{\xi}_{\mathbf{C}}\right| \geq \epsilon\right) \leq \sup_{\mathbf{C}} \frac{nE\left[\left(\boldsymbol{\xi}_{\mathbf{C}}^T A \boldsymbol{\xi}_{\mathbf{C}} - \boldsymbol{\xi}_{\mathbf{C}}^T A_d \boldsymbol{\xi}_{\mathbf{C}}\right)^2\right]}{\epsilon^2} \tag{23}$$

Since the block diagonal elements of A_d equal those of A, it sufices to show that the off diagonal blocks of A are negligible. For $i_1 \neq i_2$, every element of the block (i_1, i_2) equals $\frac{1}{n(n-1)p}$. We will show that the second moment on the right hand side of (23) conditionally on $\mathbf{Z}$ goes to zero, and therefore the unconditional second moment also does. To that end, write

$$\begin{aligned}
&\sup_{\mathbf{C}} \frac{nE\left[\left(\boldsymbol{\xi}_{\mathbf{C}}^T A \boldsymbol{\xi}_{\mathbf{C}} - \boldsymbol{\xi}_{\mathbf{C}}^T A_d \boldsymbol{\xi}_{\mathbf{C}}\right)^2 |\mathbf{Z}\right]}{\epsilon^2} \qquad (24)\\
&= n\left(\frac{1}{n(n-1)p}\right)^2 \sup_{\mathbf{C}} E\left(\sum_{i_1\neq i_2}\sum_{i_3\neq i_4}\sum_{j_1,j_2,j_3,j_4=1}^{n} \xi_{j_1}\xi_{j_2}\xi_{j_3}\xi_{j_4} I(j_k \in W_{i_k}(\mathbf{C}), k=1,\ldots,4)|\mathbf{Z}\right)\\
&= n\left(\frac{1}{n(n-1)p}\right)^2 \sup_{\mathbf{C}} \sum_{i_1\neq i_2}\sum_{i_3\neq i_4}\sum_{j_1,j_2,j_3,j_4=1}^{n} E\left(\xi_{j_1}\xi_{j_2}\xi_{j_3}\xi_{j_4}|\mathbf{Z}\right) I(j_k \in W_{i_k}(\mathbf{C}), k=1,\ldots,4)
\end{aligned}$$

The expected value in this sum is different from zero, only if $\xi_{j_1}, \ldots, \xi_{j_4}$ consists of two pairs of equal observations, or $j_1 = j_2 = j_3 = j_4$. Since there are $O(n^2p^4)$ terms for the former case to happen and $O(np^4)$ for the latter case to happen, and the magnitude of these terms is not affected by $\mathbf{C}$, the order of (24) is $O\left(\frac{n}{p}\frac{1}{n^4p^2}n^2p^4\right) = o(1)$, and this completes the proof. □

Corollary A.1. *Let $A_d = diag\{B_1, ..., B_n\}$, with $B_i = \frac{1}{n(p-1)}[\mathbf{J}_p - \mathbf{I}_p]$, $sd = \sqrt{\frac{2p(2p-1)}{3(p-1)}\tau^2}$, and $\boldsymbol{\xi}_{\mathbf{C}}$ be defined in (4) with $\mathbf{C}$ instead of $\mathbf{C}_\theta$. Then, under the assumptions of Theorem 2.1 we have*

$$\begin{aligned}
&\sup_{\mathbf{C}}\sup_{t}\left|P\left(\frac{n^{1/2}\boldsymbol{\xi}_{\mathbf{C}}^T A \boldsymbol{\xi}_{\mathbf{C}}}{sd} \leq t\right) - \Phi(t)\right| \to 0 \text{ if and only if}\\
&\sup_{\mathbf{C}}\sup_{t}\left|P\left(\frac{n^{1/2}\boldsymbol{\xi}_{\mathbf{C}}^T A_d \boldsymbol{\xi}_{\mathbf{C}}}{sd} \leq t\right) - \Phi(t)\right| \to 0.
\end{aligned}$$

Proof. Write

$$\frac{n^{1/2}\boldsymbol{\xi}_{\mathbf{C}}^T A \boldsymbol{\xi}_{\mathbf{C}}}{sd} = \frac{n^{1/2}\boldsymbol{\xi}_{\mathbf{C}}^T A_d \boldsymbol{\xi}_{\mathbf{C}}}{sd} + \frac{n^{1/2}\left(\boldsymbol{\xi}_{\mathbf{C}}^T A \boldsymbol{\xi}_{\mathbf{C}} - \boldsymbol{\xi}_{\mathbf{C}}^T A_d \boldsymbol{\xi}_{\mathbf{C}}\right)}{sd}.$$

Now, for any t

$$\begin{aligned}
&\sup_{\mathbf{C}}\left|P\left(\frac{n^{1/2}\boldsymbol{\xi}_{\mathbf{C}}^T A\boldsymbol{\xi}_{\mathbf{C}}}{sd}\le t\right)-\Phi(t)\right| \\
&=\sup_{\mathbf{C}}\left|P\left(\frac{n^{1/2}\boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}}{sd}\le t-\frac{n^{1/2}\left(\boldsymbol{\xi}_{\mathbf{C}}^T A\boldsymbol{\xi}_{\mathbf{C}}-\boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}\right)}{sd},\left|\frac{n^{1/2}\left(\boldsymbol{\xi}_{\mathbf{C}}^T A\boldsymbol{\xi}_{\mathbf{C}}-\boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}\right)}{sd}\right|\le\epsilon\right)\right. \\
&\left.+P\left(\frac{n^{1/2}\boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}}{sd}\le t-\frac{n^{1/2}\left(\boldsymbol{\xi}_{\mathbf{C}}^T A\boldsymbol{\xi}_{\mathbf{C}}-\boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}\right)}{sd},\left|\frac{n^{1/2}\left(\boldsymbol{\xi}_{\mathbf{C}}^T A\boldsymbol{\xi}_{\mathbf{C}}-\boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}\right)}{sd}\right|\ge\epsilon\right)-\Phi(t)\right| \\
&\le\sup_{\mathbf{C}}\max\left\{\left|P\left(\frac{n^{1/2}\boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}}{sd}\le t+\epsilon\right)-\Phi(t)\right|,\left|P\left(\frac{n^{1/2}\boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}}{sd}\le t-\epsilon\right)-\Phi(t)\right|\right\}\right. \\
&\qquad\left.+\sup_{\mathbf{C}}P\left(\left|\frac{n^{1/2}\left(\boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}-\boldsymbol{\xi}_{\mathbf{C}}^T A\boldsymbol{\xi}_{\mathbf{C}}\right)}{sd}\right|\ge\epsilon\right).\right. \qquad (25)
\end{aligned}$$

The last term in (25) goes to zero by Lemma A.1. Thus,

$$\begin{aligned}
&\sup_{\mathbf{C}}\sup_t\left|P\left(\frac{n^{1/2}\boldsymbol{\xi}_{\mathbf{C}}^T A\boldsymbol{\xi}_{\mathbf{C}}}{sd}\le t\right)-\Phi(t)\right| \\
&\le\sup_{\mathbf{C}}\max\left\{\sup_t\left|P\left(\frac{n^{1/2}\boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}}{sd}\le t\right)-\Phi(t-\epsilon)\right|,\sup_t\left|P\left(\frac{n^{1/2}\boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}}{sd}\le t\right)-\Phi(t+\epsilon)\right|\right\} \\
&\qquad+o(1) \\
&\le\sup_{\mathbf{C}}\sup_t\left|P\left(\frac{n^{1/2}\boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}}{sd}\le t\right)-\Phi(t)\right|+\sup_t|\Phi(t)-\Phi(t+\epsilon)|+o(1).
\end{aligned}$$

Letting $\epsilon\to 0$,

$$\lim_{n\to\infty}\sup_{\mathbf{C}}\sup_t\left|P\left(\frac{n^{1/2}\boldsymbol{\xi}_{\mathbf{C}}^T A\boldsymbol{\xi}_{\mathbf{C}}}{sd}\le t\right)-\Phi(t)\right|\le\lim_{n\to\infty}\sup_{\mathbf{C}}\sup_t\left|P\left(\frac{n^{1/2}\boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}}{sd}\le t\right)-\Phi(t)\right|.$$

Using similar steps, it can be shown that

$$\lim_{n\to\infty}\sup_{\mathbf{C}}\sup_t\left|P\left(\frac{n^{1/2}\boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}}{sd}\le t\right)-\Phi(t)\right|\le\lim_{n\to\infty}\sup_{\mathbf{C}}\sup_t\left|P\left(\frac{n^{1/2}\boldsymbol{\xi}_{\mathbf{C}}^T A\boldsymbol{\xi}_{\mathbf{C}}}{sd}\le t\right)-\Phi(t)\right|,$$

completing the proof. □

Lemma A.2. *Let $S_V(\mathbf{C})$ be defined as in (18). Under the assumptions of Theorem 2.1,*

$$\sup_{\mathbf{C}}P\left(\left|\frac{S_V(\mathbf{C})}{\sqrt{n}\ sd}\right|\ge\epsilon\right)\to 0$$

Proof. For any $\epsilon > 0$, since $V_i(\mathbf{C})$ are independent,

$$\sup_{\mathbf{C}} P\left(n^{-1/2}|\sum_{i=1}^{r_n} V_i(\mathbf{C})| \geq \epsilon\right) \leq \sup_{\mathbf{C}} \sum_{i=1}^{r_n} P\left(|V_i(\mathbf{C})| \geq \epsilon n^{1/2} r_n^{-1}\right)$$
$$\leq \sup_{\mathbf{C}} \sum_{i=1}^{r_n} \frac{E(V_i(\mathbf{C})^4)}{\epsilon^4 n^2 r_n^{-4}} \leq K\epsilon^{-4} n^{-2} r_n^5 (p^2)^2 = o(1),$$

where the last inequality follows from the fact that $E(V_i^4(\mathbf{C})) \leq K(p^2)^2$. □

Lemma A.3. *Let $S_U(\mathbf{C})$ and $S_V(\mathbf{C})$ be defined as in (18). Under the assumptions of Theorem 2.1,*

$$\sup_{\mathbf{C}} \left| P\left(\frac{S_U(\mathbf{C})}{\sqrt{n}sd} \leq t - \frac{S_V(\mathbf{C})}{\sqrt{n}sd}, \left|\frac{S_V(\mathbf{C})}{\sqrt{n}sd}\right| \leq \epsilon\right) - \Phi(t)\right| \to 0. \tag{26}$$

Proof. Note that, using the Berry Essen bound (see Shorack (Probability for Statisticians)), and the fact that $Var(\frac{S_U(\mathbf{C})}{\sqrt{n}}) \to \mathrm{sd}^2$ as shown in the proof of Theorem 2.1, we have

$$\sup_{\mathbf{C}} \sup_t \left| P\left(\frac{S_U(\mathbf{C})}{\sqrt{n}\mathrm{sd}} \leq t\right) - \Phi(t)\right| \leq 9 \sup_{\mathbf{C}} \frac{\sum_{i=1}^{r_n} E\left|U_i(\mathbf{C})\right|^3}{\left[\sum_{i=1}^{r_n} \mathrm{Var}(U_i(\mathbf{C}))\right]^{3/2}} = O\left(\frac{1}{\sqrt{r_n}}\right) = o(1). \tag{27}$$

Let $t^* = t - \frac{S_V(\mathbf{C})}{\sqrt{n}\mathrm{sd}}$, then

$$\begin{aligned}
&\sup_{\mathbf{C}} \left| P\left(\frac{S_U(\mathbf{C})}{\sqrt{n}\mathrm{sd}} \leq t - \frac{S_V(\mathbf{C})}{\sqrt{n}\mathrm{sd}}, \left|\frac{S_V(\mathbf{C})}{\sqrt{n}\mathrm{sd}}\right| \leq \epsilon\right) - \Phi(t)\right| \\
= &\sup_{\mathbf{C}} \left| P\left(\frac{S_U(\mathbf{C})}{\sqrt{n}\mathrm{sd}} \leq t^*, \left|\frac{S_V(\mathbf{C})}{\sqrt{n}\mathrm{sd}}\right| \leq \epsilon\right) - P\left(\frac{S_U(\mathbf{C})}{\sqrt{n}\mathrm{sd}} \leq t^*\right)\right. \\
&\qquad \left. + P\left(\frac{S_U(\mathbf{C})}{\sqrt{n}\mathrm{sd}} \leq t^*\right) - \Phi(t^*) + \Phi(t^*) - \Phi(t)\right| \\
\leq &\sup_{\mathbf{C}} \left| P\left(\frac{S_U(\mathbf{C})}{\sqrt{n}\mathrm{sd}} \leq t^*, \left|\frac{S_V(\mathbf{C})}{\sqrt{n}\mathrm{sd}}\right| \leq \epsilon\right) - P\left(\frac{S_U(\mathbf{C})}{\sqrt{n}\mathrm{sd}} \leq t^*\right)\right| \\
&\qquad + \sup_{\mathbf{C}} \left| P\left(\frac{S_U(\mathbf{C})}{\sqrt{n}\mathrm{sd}} \leq t^*\right) - \Phi(t^*)\right| + \left|\Phi(t^*) - \Phi(t)\right|.
\end{aligned} \tag{28}$$

The first term in (28) goes to zero by continuity of measures, since by Lemma A.2 $P\left(\left|\frac{S_V(\mathbf{C})}{\sqrt{n}\mathrm{sd}}\right| \leq \epsilon\right) \to 1$. The second term in (28) goes to zero by (27), and the third term goes to zero by the continuity of $\Phi(.)$.

□

Lemma A.4. *Let $X_1, \ldots, X_n$ be iid[F], and let $\hat{F}_n(x)$ be the corresponding empirical distribution function. Then, for any constant c,*

$$sup_{x_i,x_j}\left\{|F(x_i)-F(x_j)|I\left[|\hat{F}(x_i)-\hat{F}(x_j)|\leq\frac{c}{n}\right]\right\}=O_p\left(\frac{1}{\sqrt{n}}\right).$$

Proof. By the Dvoretzky, Kiefer and Wolfowitz (1956) theorem, we have that $\forall\epsilon\geq 0$,

$$P\left(\sup_x|\hat{F}_n(x)-F(x)|\geq\epsilon\right)\leq Ce^{-2n\epsilon^2}.$$

Therefore, $|\hat{F}(x)-F(x)|=O_p\left(\frac{1}{\sqrt{n}}\right)$ uniformly on x. Hence, writing

$$|F(x_i)-F(x_j)| \;=\; |F(x_i)-\hat{F}_n(x_i)+\hat{F}_n(x_i)-F(x_j)+\hat{F}_n(x_j)-\hat{F}_n(x_j)|,$$

it follows that $sup_{x_i,x_j}\left\{|F(x_i)-F(x_j)|I\left[|\hat{F}(x_i)-\hat{F}(x_j)|\leq c/n\right]\right\}$ is less than or equal to

$$\begin{aligned}
&sup_{x_i,x_j}\left\{|F(x_i)-\hat{F}_n(x_i)|+|\hat{F}_n(x_j)-F(x_j)|\right\}\\
&+sup_{x_i,x_j}\left\{|\hat{F}_n(x_i)-\hat{F}_n(x_j)|\right\}I\left[|\hat{F}_n(x_i)-\hat{F}_n(x_j)|\leq\frac{c}{n}\right]\\
=\;& O_p\left(\frac{1}{\sqrt{n}}\right)+O_p\left(\frac{1}{\sqrt{n}}\right)+O_p\left(\frac{1}{n}\right).
\end{aligned}$$

This completes the proof of the lemma. □

Lemma A.5. *With W_i be defined in (3), and any Lipschitz continuous function $g(x)$,*

$$\frac{1}{p}\sum_{j=1}^{n}g(x_{2j})I(j\in W_i)-g(x_{2i})=O_p\left(\frac{1}{\sqrt{n}}\right),$$

uniformly in $i=1,\ldots,n$.

Proof. First note that by the Lipschitz continuity and the Mean Value Theorem we have

$$|g(x_{2j})-g(x_{2i})|\leq M|x_{2j}-x_{2i}|\leq M|F_{X_2}(x_{2j})-F_{X_2}(x_{2i})|/f_{X_2}(\tilde{x}_{ij}),$$

for some constant M, where $\tilde{x}_{ij}$ is between x_{2j} and x_{2i}. Thus,

$$\begin{aligned}
&\left|\frac{1}{p}\sum_{j=1}^{n}g(x_{2j})I(j\in W_i)-g(x_{2i})\right|\leq\frac{1}{p}\sum_{j=1}^{n}|g(x_{2j})-g(x_{2i})|I\left[|\hat{F}_{X_2}(x_{2i})-\hat{F}_{X_2}(x_{2j})|\leq\frac{p-1}{2n}\right]\\
&\leq\frac{M}{p}\sum_{j=1}^{n}\frac{|F_{X_2}(x_{2j})-F_{X_2}(x_{2i})|}{f_{X_2}(\tilde{x}_{ij})}I\left[|\hat{F}_{X_2}(x_{2i})-\hat{F}_{X_2}(x_{2j})|\leq\frac{p-1}{2n}\right]=O_p\left(\frac{1}{\sqrt{n}}\right),
\end{aligned}$$

where the last equality follows from Lemma A.4 and the assumption that f_{X_2} remains bounded away from zero. □

References

Abramovich, F.,Benjamini, Y., Donoho, D.L. and Johnstone,I.M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, 34, 584-653.

Akritas, M. G. and Papadatos, N. (2004). Heterocedastic One-Way ANOVA and Lack-of-Fit Tests. *Journal of the American Statistical Association*, 99, Theory and Methods.

Alon, U., Barkai, N., Notterdam, D., Gish, K., Ybarra, S., Mack, D. and Levine, A. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Science* 96, 67456750.

Benjamini, Y.; Gavrilov, Y. (2009). A Simple Forward Selection Procedure Based on False Discovery Rate Control. *The Annals of Applied Statistics*, 3, 179-198.

Benjamini, Y.; Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57 (1): 289300.

Benjamini, Y., Krieger, A.M., Yekutieli, D. (2006). Adaptive Linear Step-up False Discovery Rate controlling procedures. *Biometrika*, 93 (3): 491-507.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statististics*, 29, 1165-1188.

Birge, L. and Massart, P. (2001) A generalized Cp criterion for Gaussian model. *Technical report*, Lab. De Probabilities, Univ. Paris VI. (http://www.proba.jussieu.fr/mathdoc/preprints/index.html#2001)

Bunea, F., Wegkamp, M. and Auguste, A. (2006). Consistent variable selection in high dimensional regression via multiple testing. *Journal of Statistical Planning and Inference,* 136, 4349-4364.

Candes, E., and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35, 2313-2351.

Dettling, M. and Bhlmann, P. (2002). Supervised clustering of genes.Genome Biology 3(12): research0069.1-0069.15.

Dettling, M. and Bhlmann, P. (2004). Finding predictive gene groups from microarray data. Journal of Multivariate Analysis 90, 106-131.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81, 42555.

Dvoretzky, A.; Kiefer, J. and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27 (3), 642669.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression. *The Annals of Statistics*, 32, 407-499.

Fan, J. and Jiang, J.(2005). Nonparametric Inferences for Additive Models. *Journal of the American Statistical Association*, 100, 890-907.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348-1360.

Foster, D. P. and Stine, R. A. (2004). Variable selection in data mining: building a predictive model for bankruptcy. *Journal of the American Statistical Association*, 99, 303-313.

Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable Selection in Nonparametric Additive Models. Available at http://faculty.wcas.northwestern.edu/∼jlh951/papers/HHW-npam.pdf

Li, K. C., (1991). Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association*, 86, 316-327.

Li, L., Cook, R., D. and Nachtsheim, C. (2005). Model-free variable selection. *Journal of the Royal Statistical Society - B*, 67(2), 285-299.

Li, R. and Liang, H. (2008). Variable selection in Semiparametric Regression Modeling. *The Annals of Statistics*, 36, 261-286.

Ma, S., Song, X. and Huang, J. (2007). Supervised group Lasso with applications to microarray data analysis. *BMC Bioinformatics*, 8:60.

Park, M., Y., Hastie, T. and Tibshirani, R. (2007). Averaged gene expressions for regression. *Biostatistics*, 212-227.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *The Annals of Statistics*, 12, 1215-1230.

Storlie, C. B, Bondell, H. D, Reich, B. J, Zhang, H. H. (2011). Surface Estimation, Variable Selection, and the Nonparametric Oracle Property. *Statistica Sinica*, 21(2), 679-705.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statististical Society B*, 58, 267-288.

Tibshirani, R. and Knight, K. (1999). The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society B*, 61, 529-546.

Wang, H. and Xia, Y. (2008). Shrinkage estimation of the varying coefficient model. Journal of the American Statistical Association, 104,747-757.

Wang, L., Akritas, M. G. and Keilegom, I.V. (2008). An ANOVA-type Nonparametric Diagnostic Test for Heterocedastic Regression Models. *Journal of Nonparametric Statistics*, 20, 365-382.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Association - B*, 68(1), 49-67.

Zambom, A. Z. (2012). Hypothesis testing and Variable Selection in Nonparametric Regression. *Doctoral Dissertation*, Department of Statistics, Penn State University.

Zambom, A. Z. and Akritas, M. (2012). Nonparametric model checking and variable selection. Submitted for publication. (Available on line at http://www.stat.psu.edu/∼mga/papers/Zambom/ZambomAkritasVS.pdf/.)

Zou, H. (2006). The Adaptive Lasso and its Oracle Properties. *Journal of the American Statistical Association*, 101(476), 1418-1429.